

Tsirelson's bound from a Generalised Data Processing Inequality

Oscar C. O. Dahlsten,^{1,2,3} Daniel Lercher,^{1,4} and Renato Renner¹

¹*Institute for Theoretical Physics, ETH Zürich, 8093 Zurich, Switzerland*

²*Center for Quantum Technologies, National University of Singapore, Republic of Singapore*

³*Atomic and Laser Physics, Clarendon Laboratory,*

University of Oxford, Parks Road, Oxford OX13PU, United Kingdom

⁴*Department of Mathematics, Technische Universität München, 85748 Garching, Germany*

(Dated: July 4, 2012)

The strength of quantum correlations is bounded from above by Tsirelson's bound. We establish a connection between this bound and the fact that correlations between two systems cannot increase under local operations, a property known as the *data processing inequality*. More specifically, we consider arbitrary convex probabilistic theories. These can be equipped with an entropy measure that naturally generalizes the von Neumann entropy, as shown recently in [Short and Wehner, Barnum et. al.]. We prove that if the data processing inequality holds with respect to this generalized entropy measure then the underlying theory necessarily respects Tsirelson's bound. We moreover generalise this statement to any entropy measure satisfying certain minimal requirements. A consequence of our result is that not all of the entropic relations used to derive Tsirelson's bound via information causality in [Pawlowski et. al.] are necessary.

PACS numbers: 03.65.Ta, 03.65.Ud

Introduction.—Quantum mechanics departs fundamentally from any classical theory by allowing non-local correlations [1]. The existence of such correlations has been extensively verified in experiments (up to a few loopholes), see e.g. [2]. As was shown by Bell, these correlations imply that the world is not both local and realist, two standard assumptions underpinning the classical mechanical world-view [1]. Apart from their fundamental theoretical interest, non-local correlations are also of technological importance, for example as the essential ingredient in Ekert-style quantum cryptographic schemes [3].

However there is a limit to how much local realism is violated. The strength of quantum correlations are themselves upper bounded by Tsirelson's bound [4, 5]. This is a non-trivial bound, because it is conceivable to violate Bell-inequalities more than quantum theory, without having a theory which is signalling (allows instantaneous information transfer across space). For example it is possible to conceive of *PR-boxes*, also known as *non-local boxes*, hypothetical systems which maximally violate the CHSH Bell-inequality, without being signalling [6].

The question then arises as to whether one can associate a fundamental assumption about nature other than non-signalling with Tsirelson's bound. Such an assumption could then be labelled a fundamental principle underpinning quantum theory, and possibly form part of a much-sought-for set of principles from which quantum theory could be derived.

There has already been significant effort in this direction. For example, it is now known that the existence of maximally Bell-violating correlations would lead to some communication complexity problems becoming trivial [7, 8], the possibility of oblivious transfer [9], weaker uncertainty relations [10], general invalidation of quantum theory locally [11] and severely limited dynam-



FIG. 1: The Data Processing Inequality states that the correlations between A and B cannot increase under a local operation T on B . More specifically $H(A|B) \leq H(A|T(B))$.

ics [12, 13].

A recent string of related papers have moreover been concerned with a principle called information causality [14–16]. A great advantage with this principle is that the exact Tsirelson's bound is recovered, i.e. it rules out any stronger correlations, not just the maximally strong ones. The principle amounts to placing a limit on how well two separated parties can perform in a particular game (van Dam's game [9]) where they share a resource state. This limits the resource state in such a way that Tsirelson's bound is recovered. Whilst the original interpretation of information causality as a particularly simple generalisation of non-signalling has been questioned (see e.g. [17]), the principle is —as mentioned above— powerful.

Intriguingly, in the proof that information causality holds in quantum theory, a specific limited set of information-theoretic theorems are used. One may thus replace information causality as a postulate with those information-theoretic theorems. This is attractive if one seeks an information-theoretic set of principles for quantum theory. In order to discuss the validity of such theorems outside of quantum theory, however, one needs definitions of the relevant entropies for general probabilistic theories. Fortunately, such definitions were recently proposed and investigated in [17–19]. In [17, 18] information

causality is also discussed. In [18] three sufficient conditions under which a generalised probabilistic theory respects information causality are determined. In [17] it is shown that if one follows the information causality proof in the case of *box-world*, the theory with PR-boxes and all other non-signalling distributions, the proof breaks down at the point where one needs to assume the so-called strong subadditivity of entropy. An alternative approach to deriving information causality from more basic entropic principles appears in [20]. These recent works when taken together suggest that one may hope for a small and operationally motivated set of information-theoretic relations from which Tsirelson's bound, and perhaps even quantum theory, can be derived.

We here investigate the Data Processing Inequality as such a principle. This essentially states that correlations, quantified via conditional entropies, cannot increase under local operations, see Fig. 1. In order to define this in general, we use an entropy proposed in [17], which naturally generalizes the von Neumann entropy (and reduces to the latter in the case of quantum theory). We prove that, surprisingly, this generalised Data Processing Inequality alone implies Tsirelson's bound.

We proceed as follows. We firstly describe the framework of generalised probabilistic theories within which we work. Then we define Tsirelson's bound as well as information causality. We go on to describe how to define entropy in an operational manner as in [17]. This is used to define the generalised Data Processing Inequality (DPI). We then prove that DPI implies Tsirelson's bound. This involves proving a more general theorem of which the main result is a corollary. Finally we compare the results to previous ones and discuss the implications and interpretation of the principle.

Convex, operational, probabilistic theories.—

We use the framework of convex probabilistic theories [12, 21, 22]. This amounts to taking the minimalistic pragmatic view that the operational content of a theory is in the predicted statistics of measurement outcomes.

The *state* of a system by definition determines the probabilities of all possible measurement outcomes. The state is completely specified, again by definition, by the probabilities for the outcomes of k so-called *fiducial* measurements $0, \dots, k-1$. k may be significantly smaller than the total number of measurements (e.g. in quantum theory there is a continuum of measurements but $k = d^2$ for a state on a Hilbert space of dimension d). If these fiducial measurements each have l possible outcomes $0, \dots, l-1$ we will say that the system is of type (k, l) .

We can thus write a (normalised) state as a list of $P(i|j)$, denoting the probability of getting outcome i if fiducial measurement j is performed. We represent this by \vec{P} . The normalisation of the state is $|\vec{P}| := \sum_i P(i|j)$ and is for all valid states independent of the choice of fiducial measurement j . A state is said to be normalised if $|\vec{P}| = 1$ and subnormalised if $|\vec{P}| < 1$.

We assume that the set of allowed normalized states \mathcal{S}

is closed and convex (so that any probabilistic mixture of states is an allowed state). We say a state is *pure* if it cannot be written as a convex mixture of other states. A *theory* is defined by the set of allowed states, \mathcal{S} , as well as the set of allowed transformations.

Transformations take states to states. They must be linear as probabilistic mixtures of different states must be conserved [12]. Transformations can thus be modelled as $\vec{P} \mapsto M \cdot \vec{P}$, where M is a matrix. If one performs a measurement with several outcomes each outcome is associated with a certain transform M_i . The unnormalized state associated with the i -th outcome is $M_i \cdot \vec{P}$, and the associated probability of the i -th outcome is given by the normalisation factor after the transformation: $|M_i \cdot \vec{P}|$.

If one is only interested in the probabilities of the different outcomes of a measurement one can always associate with a transformation $\{M_i\}$ a set of vectors $\{R_i\}$ such that $\vec{R}_i \cdot \vec{P} = |M_i \cdot \vec{P}| \forall \vec{P} \in \mathcal{S}$. Consequently, for a normalized state \vec{P} , $\vec{R}_i \cdot \vec{P}$ is the probability of the i th outcome.

It is also possible to combine single systems to form multipartite systems. If one performs local operations on the systems A and B the final unnormalized state of the joint system does by assumption not depend on the temporal ordering of the operations. A direct consequence of this is the no-signaling principle: measuring system B cannot give information about what transformation was applied on A [12].

We will make the non-trivial but standard assumption that the global state of a bipartite system can be completely determined by specifying joint probabilities of outcomes for fiducial measurements performed simultaneously on each subsystem. Accordingly, the joint state of two parties is uniquely specified by the list $P(ii'|jj')$, denoting the probability of getting the outcomes i and i' if one performs fiducial measurement j on A and j' on B .

For a joint state \vec{P}_{AB} , the *marginal* (also called *reduced*) state of system A , denoted \vec{P}_A , is given by $P_A(i|j) \equiv \sum_{i'} P_{AB}(ii'|jj')$. Similarly, the *conditional marginal* state $\vec{P}_{A|B:k,l}$ is defined by

$$P_{A|B:k,l}(i|j) \equiv \frac{P_{AB}(ik|jl)}{P_B(k|l)}. \quad (1)$$

This represents the state of system A after a fiducial measurement l was performed on system B and the outcome k was obtained.

It was shown in [12] that denoting the vector spaces containing the vectors \vec{P}_{AB} , \vec{P}_A , and \vec{P}_B by V_{AB} , V_A and V_B , respectively, one can relate the spaces by $V_{AB} = V_A \otimes V_B$ (\otimes being the tensor product). One assumes that for $\vec{P}_A \in \mathcal{S}_A$ and $\vec{P}_B \in \mathcal{S}_B$ we have $\vec{P}_A \otimes \vec{P}_B \in \mathcal{S}_{AB}$. This implies that any $\vec{P}_{AB} \in \mathcal{S}_{AB}$ can be written as $\vec{P}_{AB} = \sum_i r_i \vec{P}_A^i \otimes \vec{P}_B^i$ with $\vec{P}_A^i \in \mathcal{S}_A$ and $\vec{P}_B^i \in \mathcal{S}_B$ normalized and pure and $r_i \in \mathbb{R}$ [12].

For a transformation on system A defined by $\vec{P}_A \mapsto \vec{P}_{A'} = M_A \cdot \vec{P}_A$ the transformation of the joint system

is given by $\vec{P}_{AB} \mapsto \vec{P}_{A'B} = (M_A \otimes \mathbf{1}) \cdot \vec{P}_{AB}$ [12]. We demand that transformations M_A on any system A are *well-defined*, meaning $(M_A \otimes I_B) \cdot \vec{P}_{AB} \in \mathcal{S}_{AB}$ whenever $\vec{P}_{AB} \in \mathcal{S}_{AB}$ for all types of system B .

In the following, we will always assume that the set of transformations allowed by the theory includes removing systems (which corresponds to taking the marginal state, as defined above) and adding a system, taking $\vec{P}_A \mapsto \vec{P}_A \otimes \vec{P}_B$.

We also demand that the theory contains ‘classical’ systems of type $(1, d)$ for all $d \in \mathbb{N}$. We call the trivial classical system of type $(1, 1)$ the vacuum (V). We shall in our proofs, taking inspiration from [21], use the fact that the state of a classical system can be *cloned*—see the technical supplement.

As shown e.g. in [22], finite dimensional quantum theory as well as classical probability theory fit into this framework. So does *box-world* [12]. This allows all states on discrete sets of measurements that are non-signalling. The simplest non-trivial example of this is for elementary systems of type $(2, 2)$. The joint state space of two such systems includes PR-boxes. A key difference between box-world and quantum theory is that only the latter respects Tsirelson’s bound.

Tsirelson’s bound.—The quantum correlation strength as quantified by the CHSH Bell inequality [23] is upper bounded by Tsirelson’s bound [4, 5].

Definition 1 (Tsirelson’s bound). *Consider two systems A and B , with two choices of measurements (0 or 1) and two outputs each (a and b). Define the quantity*

$$S := p(a = b|00) + p(a = b|01) + p(a = b|10) + p(a \neq b|11).$$

The theory governing the systems is said to satisfy Tsirelson’s bound if $2 - \sqrt{2} \leq S \leq 2 + \sqrt{2}$ for any states allowed by the theory.

A PR-box (also known as a non-local box) is designed to have $S=0$ or 4 , thus maximally violating the Tsirelson bound [6]. It is defined (up to relabellings of measurement choices and outcomes) to be a state where

$$p(a = b|00) = p(a = b|01) = p(a = b|10) = p(a \neq b|11) = 1$$

and the local marginal states are uniformly random.

Information causality.—Let there be two space-like separated parties, Alice and Bob which share an arbitrary no-signaling resource. Alice then receives a random bit-string $\vec{a} = (a_0, \dots, a_{N-1})$, which is not known to Bob. The bits a_i are unbiased and independently distributed. At the same time Bob gets a random variable $b \in \{0, \dots, N-1\}$, which is unknown to Alice. Alice is free to make use of her local resources in order to prepare a classical bit-string \vec{x} of length m which she sends to Bob. Bob, having received Alice’s message, is then asked to guess the value of a_b as best as he can. Let us denote Bob’s guess by β . The efficiency of Alice’s and Bob’s strategy can be quantified by $I \equiv \sum_i I_{\text{Sh}}(a_i : \beta|b = i)$

where $I_{\text{Sh}}(a_i : \beta|b = i)$ is the Shannon mutual information between a_i and β , computed under the condition that Bob has received $b = i$.

Definition 2 (Information Causality). *A theory is said to respect information causality if in the above game $I \leq m$ for any allowed resource state.*

It was shown in [14] that information causality implies Tsirelson’s bound.

General entropy definition.—We now recount certain results from recent research into how to quantify entropy in general probabilistic theories [17–19]. We shall in particular use a definition of entropy for general theories from [17] which is based on the Shannon entropy. This is highly analogous to how the von Neumann entropy generalises the Shannon entropy $H_{\text{Sh}}(\vec{P}) = -\sum_i P_i \log P_i$ to the quantum case. The intuition is that the von Neumann entropy is the minimal Shannon entropy over all measurements. Actually it is over all *fine-grained* measurements (explained below).

Note that one can in general define the Shannon entropy associated with a measurement e as $H_{\text{Sh}}(e(\vec{P})) = -\sum_i (\vec{R}_i^e \cdot \vec{P}) \log(\vec{R}_i^e \cdot \vec{P})$

Definition 3 (Entropy [17]). *For every normalized state $\vec{P} \in \mathcal{S}$ the entropy $H(\vec{P})$ is given by*

$$H(\vec{P}) \equiv \inf_{e \in \mathcal{M}^*} H_{\text{Sh}}(e(\vec{P})). \quad (2)$$

$e(\vec{P})$ denotes the classical probability distribution for the different outcomes of e and the minimization is over the set of all fine-grained measurements \mathcal{M}^* .

\mathcal{M}^* above is defined to be the set of measurements which have no *non-trivial fine-grainings*. A *fine-graining* is a subdivision of one outcome into several different outcomes. A *trivial* fine-graining is one where the resulting outcomes do not have independent probabilities, or more formally, where the vectors representing the respective effects are proportional to the effect-vector associated with the original coarse-grained outcome.

The restriction to minimizing over \mathcal{M}^* is important. If one allowed coarse-grained measurements the entropy could always be reduced arbitrarily by grouping outcomes together into single outcomes. It is natural to draw the line at trivial fine-grainings since no more information is yielded by them.

The entropy $H(\vec{P})$ can be interpreted as the minimal uncertainty that is associated with the outcome of a maximally informative measurement. It has some appealing properties: (i) H reduces to the Shannon entropy for classical probability theory and the von Neumann entropy in quantum theory, (ii) Suppose that the minimal number of outcomes for a fine-grained measurement in \mathcal{M}^* is d . Then for all states $\vec{P} \in \mathcal{S}$, $\log(d) \geq H(\vec{P}) \geq 0$ and (iii) for any $\vec{P}_1, \vec{P}_2 \in \mathcal{S}$ and any mixed state $\vec{P}_{\text{mix}} = p\vec{P}_1 + (1-p)\vec{P}_2 \in \mathcal{S}$: $H(\vec{P}_{\text{mix}}) \geq pH(\vec{P}_1) + (1-p)H(\vec{P}_2)$ [17].

For a state \vec{P}_{AB} of a bipartite system AB one defines the conditional entropy of A conditioned on B by [17]

$$H(A|B)_{\vec{P}_{AB}} \equiv H(\vec{P}_{AB}) - H(\vec{P}_B), \quad (3)$$

with \vec{P}_B the reduced state of \vec{P}_{AB} . If there are no ambiguities we drop the indices and we write $H(A)$ instead of $H(\vec{P}_A)$ and $H(AB)$ instead of $H(\vec{P}_{AB})$, and so on.

Some properties that are satisfied in quantum theory (where this entropy reduces to the von Neumann entropy) are not necessarily satisfied for arbitrary theories. In box-world, for example so-called strong subadditivity can be violated, as well as the subadditivity of the conditional entropy [17].

Data processing inequality.—The data processing inequality (DPI) is a crucial property of entropy measures which is frequently used in proofs in classical as well as quantum information theory [24, 25]. DPI quantifies the notion that local operations cannot increase correlations. A standard formulation for the classical case is that $H(X|Y) \leq H(X|g(Y))$, where X and Y are random variables which may be correlated, $H(X|Y) := H(XY) - H(Y)$, and $g(Y)$ is a function of Y only. The quantum DPI is the same, but with H denoting the von Neumann entropy.

We will here use the following generalised definition of DPI due to Short and Wehner [17].

Definition 4 (Data Processing Inequality (DPI)). *Consider two systems A and B . The data processing inequality is that for any allowed state $\vec{P}_{AB} \in \mathcal{S}_{AB}$ and for any allowed local transformation $T : \vec{P}_B \rightarrow \vec{P}'_B$*

$$H(A|B)_{\vec{P}_{AB}} \leq H(A|B')_{(\mathbf{1} \otimes T)\vec{P}_{AB}}, \quad (4)$$

where $H(\cdot|\cdot)$ denotes the conditional entropy of Eqn. (3).

Main result.—Our main result links the data processing inequality with Tsirelson's bound.

Theorem 1. *In any general probabilistic theory where the Data Processing Inequality is respected, the Tsirelson bound is respected.*

Proof. We here sketch the proof—see the appendix for the details.

We use the fact that the entropy of Def. 3 satisfies two properties: (i) $H(A|B) := H(AB) - H(B)$ (we call this COND), and (ii) it reduces to the Shannon entropy for classical systems (we call this SHAN).

We prove that for *any* theory and entropy measure H jointly satisfying COND, SHAN and DPI, Tsirelson's bound holds (where DPI has been defined using H). This implies the main theorem.

The three conditions are not trivially applicable to restrict the resource state in van Dam's game so we use them, within the framework of probabilistic theories, to derive certain more directly applicable lemmas, including: (i) $\sum_i H(A_i|\gamma) \geq H(A|\gamma)$, where A_i denotes the

i-th party of a multi-party system A (ii) $H(A) \geq H(A|B)$ with equality for product states, and (iii) for classical systems X , $H(X|Y) \geq 0$. With these lemmas and some additional arguments we show information causality is respected, and thus, by [14], Tsirelson's bound. \square

Discussion.—We have shown that the generalised DPI implies Tsirelson's bound. This addresses a question raised in [17], namely in what manner enforcing generalised entropic relations restricts the set of possible theories. It also contributes to our understanding of why Bell-violations in quantum theory respect Tsirelson's bound.

As indicated in the proof sketch, our quantitative results can be applied to more general entropy measures. In particular, for *any* entropy measure H and theory jointly satisfying COND, SHAN and DPI, we show that Tsirelson's bound holds. Thus one could alternatively have used for example the *decomposition entropy* of [17] in the statement of the main theorem as it satisfies SHAN and is defined to satisfy COND [17]. At the same time one may argue that whilst an operationally appealing definition of conditional entropy should automatically satisfy SHAN and DPI it is not clear why it should in general satisfy COND. COND may then be viewed as a *restriction* on states rather than a *definition* of conditional entropy.

One can compare our three sufficient conditions COND, SHAN and DPI to those used in [14] and [18] respectively. The entropic relations used in [14] to derive information causality were formulated in terms of a conditional mutual information $I(A : B|C)$. (It is assumed this can be defined in a more general setting, but no definition is given.) The conditions are that $I(A : B|C)$ should: be symmetric under change of A and B , be non-negative ($I \geq 0$), reduce to the Shannon mutual information for classical systems, obey the Data Processing Inequality as formulated for mutual information, and obey the chain rule $I(A : B|C) = I(A : BC) - I(A : C)$. Arguably our three relations are more minimalistic and natural than those. Moreover we show the arguments apply to particular concrete definitions of entropy and that for at least two particular definitions of conditional entropy DPI alone suffices. Consider secondly [18]. There concrete entropy definitions are proposed and studied. The definitions are very similar to [17] though the framework is not a priori exactly identical. They define three properties in terms of conditional entropy as $H(AB) - H(B)$, with H the measurement entropy: (i) 'monoentropicity' (two particular different entropy measures always have the same value), (ii) a version of the Holevo bound, and (iii) 'strong sub-additivity' (defined below). They show that those conditions imply information causality. They moreover note that conditions (ii) and (iii) can be derived from DPI defined in terms of the above conditional (measurement) entropy (more correctly they define it using mutual information $I(A : B) := H(A) + H(B) - H(AB)$ but this is equivalent in this case). Assumption (i) is

used to obtain what we here derive as Eq. A10. Thus it appears one may alternatively summarise their result on information causality as follows: DPI (in terms of COND and measurement entropy) plus mono-entropicity implies information causality. This can be compared to our Theorem 1; it is not so clear how to compare it to our more general Theorem 2, as the latter does not refer to a specific entropy measure, but to any state space and conditional entropy measure jointly satisfying DPI, COND and SHAN.

DPI is related to a condition known as *strong subadditivity* (SSA) which states that $H(A|CD) \leq H(A|C)$. SSA is *implied* by DPI since forgetting D is an allowed local operation. In the quantum case SSA also implies DPI, but this does not necessarily hold in other theories as the standard quantum proof relies on the specific quantum feature known as Stinespring dilation. In the extreme case of box-world it was already known that SSA (and thus also DPI) is violated [17]. As an example consider two classical bits x^0, x^1 and a *gbit* Z . The latter is a (2,2) system which can take any allowed distribu-

tions, i.e. its state space is the convex hull of four states wherein the two outcomes take defined values for each measurement. The classical bits are uniformly random but the gbit contains their values. Then $H(x^0|x^1Z) = 1$ whereas $H(x^0|Z) = 0$, violating SSA [17].

It is an open question whether there are theories which satisfy DPI but have states not contained in quantum theory, since Tsirelson's $2 + \sqrt{2}$ bound is insufficient to rule out all non-quantum states. Understanding this and with what DPI needs to be supplemented in order to derive quantum theory fully are natural next steps.

Acknowledgements.—We acknowledge comments on an earlier draft from J. Oppenheim, A. Short and S. Wehner, advice on references by V. Scarani, as well as funding from the Swiss National Science Foundation (grant No. 200020-135048) and the European Research Council (grant No. 258932). The research was carried out in connection with DL's Master's thesis at ETH Zurich.

Additional Note.— Similar results have been obtained independently in [26] by Al-Safi and Short.

-
- [1] J. S. Bell, *Physics* **1**, 195 (1964).
 - [2] A. Aspect, *Nature* **398**, 189 (1999).
 - [3] A. Ekert, *Phys. Rev. Lett.* **67**, 661 (1991).
 - [4] B. S. Tsirelson, *Hadronic Journal Suppl.* **8:4**, 329 (1993).
 - [5] B. S. Cirel'son, *Lett. Math. Phys.* **4**, 93 (1980).
 - [6] S. Popescu and D. Rohrlich, *Found. Phys.* **24**, 379 (1994).
 - [7] H. Buhrman, M. Christandl, F. Unger, S. Wehner, and A. Winter, *Proc. R. Soc. London, Ser. A* **462**, 1919 (2006).
 - [8] G. Brassard *et al.*, *Phys. Rev. Lett.* **96**, 250401 (2006).
 - [9] W. van Dam, (2005), arXiv:quant-ph/0501159.
 - [10] J. Oppenheim and S. Wehner, *Science* **330**, 1072 (2010), arXiv:1004.2507.
 - [11] H. Barnum, S. Beigi, S. Boixo, M. B. Elliott, and S. Wehner, *Phys. Rev. Lett.* **104**, 140401 (2010).
 - [12] J. Barrett, *Phys. Rev. A* **75**, 032304 (2007).
 - [13] D. Gross, M. Müller, R. Colbeck, and O. C. O. Dahlsten, *Phys. Rev. Lett.* **104**, 080402 (2010).
 - [14] M. Pawłowski *et al.*, *Nature (London)* **461**, 1101 (2009).
 - [15] J. Allcock, N. Brunner, M. Pawłowski, and V. Scarani, *Phys. Rev. A* **80**, 040103 (2009).
 - [16] D. Cavalcanti, A. Salles, and V. Scarani, *Nature Communications* **1** (2010).
 - [17] A. J. Short and S. Wehner, *New J. Phys.* **12**, 033023 (2010).
 - [18] H. Barnum *et al.*, *New J. Phys.* **12**, 033024 (2010).
 - [19] G. Kimura, K. Nuida, and H. Imai, *Reports on Mathematical Physics* **66**, 175 (2010).
 - [20] L.-Y. Hsu, I.-C. Yu, and F.-L. Lin, *Phys. Rev. A* **84**, 042319 (2011).
 - [21] H. Barnum, J. Barrett, M. Leifer, and A. Wilce, *Phys. Rev. Lett.* **99**, 240501 (2007).
 - [22] L. Hardy, (2001), arXiv:quant-ph/0101012.
 - [23] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, *Phys. Rev. Lett.* **23**, 880 (1969).
 - [24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 99th ed. (Wiley-Interscience, 1991).
 - [25] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, 1 ed. (CUP, 2000).
 - [26] S. Al-Safi and A. Short, (2011), arXiv:1107.4031.

Appendix A: Proof of main theorem

The main theorem is a direct corollary of a more general theorem, Theorem 2, which we state and prove in this section. Crucially, Theorem 2 does not refer to a specific entropy measure such as the measurement entropy defined above.

We require three definitions to state this theorem.

Firstly we redefine DPI, now defined without reference to a specific entropy definition.

Definition 5 (Data Processing Inequality (DPI)). *Consider two systems A and B . The data processing inequality is that for any allowed state $\vec{P}_{AB} \in \mathcal{S}_{AB}$ and for any allowed local transformation $T : \vec{P}_B \rightarrow \vec{P}'_B$*

$$H(A|B)_{\vec{P}_{AB}} \leq H(A|B')_{(1 \otimes T)\vec{P}_{AB}}. \quad (\text{A1})$$

Definition 6 (Conditional entropy (COND)). *The conditional entropy $H(A|B)$, however it is defined, must for all allowed states on AB satisfy*

$$H(A|B) = H(AB) - H(B). \quad (\text{A2})$$

Definition 7 (Reduction to Shannon entropy (SHAN)). *The entropy H must reduce to the Shannon entropy for classical systems.*

Our statements are restricted to the generalised probabilistic framework, as described in the introduction to the paper. We shall be making use of two non-trivial but operationally well-motivated types of transformations associated with that framework: *adding* and *removing* systems. An (independent) system in state \vec{P}_B is *added* by the map taking any \vec{P}_A to $\vec{P}_A \otimes \vec{P}_B$. A system is *removed* by taking the marginal distribution on the other system(s), as described in the introduction. We shall make use of the fact that this map acts to take the removed system B to the vacuum system V . The only normalised state of the vacuum is $\vec{1}_V = 1$ (this can be seen from the equivalent definition of the marginal state used e.g. in [21]). Thus, and this is another equation we shall find useful, $\vec{P}_A \otimes \vec{1}_V = \vec{P}_A \forall \vec{P}_A$.

We shall also be assuming that the entropy measure is operational, i.e. is uniquely determined by the statistics of the experiment under consideration. Thus it is for a given set-up determined by the state of the systems under consideration. More subtly, H moreover cannot depend on the order in which the state-spaces of the subsystems are composed, as this order is arbitrary; different observers describing the same experiment can make different choices here. Thus $H(AB)$ must be invariant under the interchange of systems A and B .

We are now ready to state the theorem:

Theorem 2. *For any probabilistic theory and entropy measure H satisfying COND, SHAN and DPI, Tsirelson's bound holds.*

Before proving Theorem 2 we note that the main theorem (Theorem 1) is directly implied by this statement as the entropy H referred to there satisfies COND and SHAN.

Before proving Theorem 2 we prove some lemmas which we shall need and which may be of interest in themselves.

Lemma 3. *COND and DPI imply the relation*

$$\sum_i H(A_i|\gamma) \geq H(A_1 \dots A_n|\gamma) \quad (\text{A3})$$

for any $\vec{P}_{A_1 \dots A_n} \in \mathcal{S}_{A_1 \dots A_n}$, where A_i denotes the i -th party of the total system $A_1 \dots A_n$.

Proof. Consider firstly $n = 2$. By COND we have

$$H(A_1|\gamma) + H(A_2|\gamma) - H(A_1 A_2|\gamma) = -H(A_2|A_1\gamma) + H(A_2|\gamma).$$

By DPI this is greater than or equal to 0.

To generalise the argument to $n > 2$, let A_2 be replaced by $A_2 \dots A_n$ in the previous equation. Then by the same argument

$$H(A_1|\gamma) + H(A_2 \dots A_n|\gamma) - H(A_1 A_2 \dots A_n|\gamma) \geq 0.$$

Now we can apply the previous argument to the term $H(A_2 \dots A_n|\gamma)$ to get

$$H(A_2|\gamma) + H(A_3 \dots A_n|\gamma) \geq H(A_2 \dots A_n|\gamma).$$

This process is then repeated iteratively to recover $\sum_i H(A_i|\gamma) \geq H(A_1 \dots A_n|\gamma)$. \square

Lemma 4. *For product states $\vec{P}_A \otimes \vec{P}_B$, COND, SHAN and DPI imply the relation*

$$H(A|B) = H(A). \quad (\text{A4})$$

Proof. We firstly use COND and SHAN to show that $H(A|V) = H(A)$ for any system A . This follows from the following:

$$H(A|V) = H(AV) - H(V) \quad (\text{A5})$$

$$= H(A) - 0 \quad (\text{A6})$$

$$= H(A). \quad (\text{A7})$$

(Here COND implies the first line. As V is classical and with only one measurement outcome, SHAN implies $H(V) = 0$; $\vec{P}_{AV} = \vec{P}_A$ as mentioned in the beginning of the appendix.)

We now prove the equality of the lemma by separately proving the two corresponding inequalities in both directions. Note firstly that

$$H(A) \geq H(A|B) \quad (\text{A8})$$

for any state. To see this, consider the transformation T that takes B to the vacuum system (i.e. the transformation that *removes* B as described in the introduction to the appendix). Then, using DPI,

$$H(A|B) \leq H(A|T(B)) = H(A|V) = H(A).$$

Consider secondly the inequality in the other direction, restricting ourselves to the case of product states only:

$$H(A)_{\vec{P}_A} \leq H(A|B)_{\vec{P}_A \otimes \vec{P}_B}. \quad (\text{A9})$$

This is true because $H(A) = H(A|V)_{\vec{P}_A \otimes \vec{I}_V} \leq H(A|B)_{\vec{P}_A \otimes \vec{P}_B}$, where the last step uses DPI for the transformation that creates \vec{P}_B from the vacuum state (i.e. the transformation that *adds* B as described in the introduction to the appendix).

Combining Eqns. A8 and A9 proves the claim. \square

Lemma 5. *DPI, SHAN and COND imply that for all classical systems X ,*

$$H(X|Y) \geq 0. \quad (\text{A10})$$

Proof. To prove the lemma via DPI we shall use the fact that the extremal states of classical systems can be *cloned* [21]. More specifically, we shall make use of the fact that for a classical system X_A in state $\vec{P}_A = \sum_i p_i \vec{\mu}_i$, where the $\vec{\mu}$ are pure, and another classical system X_B of the same dimensionality in any given independent pure state $\vec{\mu}_k$, there exists a map T_C such that $T_C(\vec{P}_A \otimes \vec{P}_B) = \sum_i p_i \vec{\mu}_i \otimes \vec{\mu}_i$.

We shall consider a three-party system YX_AX_B , where Y is the only non-classical sub-system. The idea is that given an arbitrary state on $YX := YX_A$, we can always bring in another independent subsystem X_B and perform a cloning operation so that X_B becomes a copy of X_A . We may then apply DPI on the cloning transformation T_C applied on X_A and X_B . We call the states before and after the cloning $\vec{P}_{YX_AX_B}^i$ and $\vec{P}_{YX_AX_B}^f$ respectively.

By DPI we then have

$$H(Y|X_AX_B)_{\vec{P}_{YX_AX_B}^i} \leq H(Y|X_AX_B)_{\vec{P}_{YX_AX_B}^f} \quad (\text{A11})$$

Note now that the left-hand-side can be simplified. COND together with Eq. (A4) imply that $H(AB) = H(A) + H(B)$ for independently prepared A and B . This can be applied here because X_B is initially in an independent state, yielding

$$H(Y|X_AX_B)_{\vec{P}_{YX_AX_B}^i} = H(Y|X_A)_{\vec{P}_{YX_AX_B}^i}.$$

Accordingly

$$H(Y|X_A)_{\vec{P}_{YX_AX_B}^i} \leq H(Y|X_AX_B)_{\vec{P}_{YX_AX_B}^f}.$$

We also note that the marginal state on YX_A is unchanged by the cloning, i.e. $\vec{P}_{YX_A}^i = \vec{P}_{YX_A}^f$, so we may for simplicity write that for the state *after* the cloning,

$$H(Y|X_A) \leq H(Y|X_AX_B). \quad (\text{A12})$$

In the following, unless stated otherwise, we consider the state after the cloning only.

Applying Eq. (A2), i.e. COND, to Eq. (A12) and undertaking some rearrangements yields

$$H(X_B|YX_A) \geq H(X_B|X_A).$$

Moreover, SHAN implies that $H(X_B|X_A) = 0$. Thus

$$H(X_B|YX_A) \geq 0.$$

Note that since X_A and X_B are operationally indistinguishable after the cloning, $H(X_A|YX_B) = H(X_B|YX_A)$. Thus we have

$$H(X_A|YX_B) \geq 0. \quad (\text{A13})$$

By DPI

$$H(X_A|Y) \geq H(X_A|YX_B). \quad (\text{A14})$$

Thus, still *after* the cloning, we have that

$$H(X_A|Y) \geq 0. \quad (\text{A15})$$

But since the state of $X_A Y$ is unchanged by the cloning transformation, this implies that the equation holds also for the (arbitrary) initial state of $X_A Y$. Recall that we used X_A to label the classical system X . We have thus shown that $H(X|Y) \geq 0$ for an arbitrary initial state on XY . \square

Lemma 6. *COND, SHAN and DPI imply the relation*

$$H(\vec{a}|B\vec{x}) \geq n - m, \quad (\text{A16})$$

where the quantities are as defined in the information causality game (\vec{a} is the classical n -bit string given to Alice, B is the non-classical resource and \vec{x} is the classical m -bit message sent to Bob).

Proof.

$$\begin{aligned} H(\vec{a}|B\vec{x}) - H(\vec{x}|\vec{a}B) &= -H(B\vec{x}) + H(\vec{a}B) \\ &= -H(B\vec{x}) + H(\vec{a}) + H(B) \\ &= H(\vec{a}) - H(\vec{x}|B) \\ &\geq H(\vec{a}) - H(\vec{x}) \\ &= n - H(\vec{x}) \\ &\geq n - m. \end{aligned}$$

The first line follows from COND. The second line is due to the combination of Eq. (A4) and Eq. (A2) and recalling that \vec{a} and B are independent. The third line uses Eq. (A2) again. The fourth line follows from Eq. (A8). The fifth and sixth lines follow from the definition of the game as well as elementary properties of the Shannon entropy, which can be exploited due to SHAN.

It follows by applying Eq. (A10) to the left hand side that $H(\vec{a}|B\vec{x}) \geq n - m$. \square

We now put together the pieces to prove Theorem 2:

Proof of Theorem 2. By lemma 6 above, we have

$$H(\vec{a}|B\vec{x}) \geq n - m.$$

By lemma 3 this implies

$$\sum_i H(a_i|B\vec{x}) \geq n - m.$$

By DPI we accordingly have that for Bob's guess $\beta = \beta(B, \vec{x}, i)$

Recall that information causality implies Tsirelson's bound.

$$\sum_i H(a_i | \beta(i)) \geq n - m,$$

□

where, by SHAN and the fact that a_i and $\beta(i)$ are both classical, H refers to the Shannon entropy.

This implies information causality, as $I_{\text{Sh}}(a_i : \beta(i)) = H(a_i) - H(a_i | \beta(i))$, so

$$\begin{aligned} \sum_i I_{\text{Sh}}(a_i : \beta(i)) &= \sum_i H(a_i) - H(a_i | \beta(i)) \\ &= n - \sum_i H(a_i | \beta(i)) \\ &\leq m. \end{aligned}$$